1

DOCUMENTATION OF DECISION-AIDING SOFTWARE:

SCORING RULE FUNCTIONAL DESCRIPTION

DECISIONS AND DESIGNS INC.

Dorothy M. Amey
Phillip H. Feuerwerger
Roy M. Gulick

July 1979

N00014-79-C-0069

ADVANCED ARPA
DECISION TECHNOLOGY
PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

1

DOCUMENTATION OF DECISION-AIDING SOFTWARE:

SCORING RULE FUNCTIONAL DESCRIPTION

by

Dorothy M. Amey, Phillip H. Feuerwerger, and Roy M. Gulick

Sponsored by

Defense Advanced Research Projects Agency
ARPA Order 3469

July 1979

DECISIONS and DESIGNS, INC.
Suite 600, 8400 Westpark Drive
P.O.Box 907
McLean, Virginia 22101
(703) 821-2828

CONTENTS

CONTENTS (Continued)

FIGURES

SCORING RULE FUNCTIONAL DESCRIPTION

1.0 INTRODUCTION

1.1 Purpose of the Functional Description

This Functional Description provides a technical delineation of the specific functions that Scoring Rule must perform. It serves as the formal basis for mutual understanding between the functional designer of the system and the software development personnel. Together with the Scoring Rule Systems Specification, it serves as the basic documentation for systems development and implementation.

1.2 References

1.2.1 Gulick, Roy M. Documentation of Decision-Aiding Software: Introductory Guide. Technical Report TR 79-1-93. McLean, Virginia: Decisions and Designs, Inc., in press.

1.2.2 Brown, Rex V.; Kahr, Andrew S.; Peterson, Cameron R. Decision Analysis for the Manager. New York: Holt, Rinehart and Winston, 1974.

1.2.3 Amey, Dorothy M.; Feuerwerger, Phillip H.; Gulick, Roy M. Documentation of Decision-Aiding Software: Scoring Rule Systems Specification. McLean, Virginia: Decisions and Designs, Inc., July 1979.

1.2.4 Amey, Dorothy M.; Feuerwerger, Phillip H.; Gulick, Roy M. Documentation of Decision-Aiding Software: Scoring Rule Users Manual. McLean, Virginia: Decisions and Designs, Inc., July 1979.

1.3 Terms and Abbreviations

1.3.1 Scoring Rule - Scoring Rule, the name of the system, is a short description of the function performed by the software, reflecting the system's method for testing, scoring, and training probability assessors.

1.3.2 SCORE - SCORE, an abbreviation for Scoring Rule, is used throughout this report to refer to the system.

1.3.3 Terms - Standard mathematical notations and decision-analytic terminology are used throughout this Functional Description. Chapter 31 of reference 1.2.2 provides additional background and insight into the basic concepts underlying the procedures implemented by SCORE.

2.0 SYSTEM SUMMARY

2.1 System Description

SCORE is a probability assessment testing, scoring, and training system. Its general purpose is to aid decision makers and intelligence analysts by improving the probability assessments which enter into decision-analytic models of complex decision problems.

The SCORE testing methodology consists of a series of questions, each having two possible answers. Several different sets of questions can be made available, and each set may itself be divided into distinct parts to be used on different testing occasions.

The overall goal of SCORE is to improve the calibration of probability assessors so as to ensure that the probability assessments that they specify truly reflect their considered beliefs. Achievement of that goal will facilitate the decision maker's making a choice consistent with the expressed beliefs about the likelihood of future events that will affect the eventual decision outcome. For a complete description of the purpose and use of SCORE, see Documentation of Decision-Aiding Software: Scoring Rule Users Manual, reference 1.2.4.

2.2 Design Objectives

The system is designed to be used interactively by end users who are relatively unsophisticated with respect to computer technology. Accordingly, the design satisfies two human-factors objectives: SCORE gives directions specifically explaining how to respond to its questions, and SCORE is generally forgiving of procedural errors by the user.

In addition, to facilitate the production of the program specification and coding necessary to implement SCORE at a physical site, the system is designed in a hierarchically structured and modular fashion. The logical structure of SCORE is explained in Documentation of Decision-Aiding Software: Scoring Rule Systems Specification, reference 1.2.3.

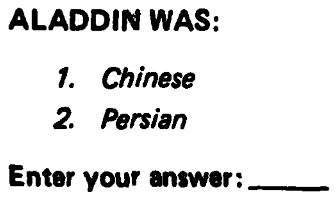
3.0 DETAILED CHARACTERISTICS

The fundamental product of SCORE is the result of a probability assessment calibration test. The SCORE system enables the user to obtain such results using a variety of assessment question sets. For example, probability assessors may be scored on general information or on highly specific subject areas.

All of the specific functions that SCORE performs are related to the probability assessment testing and scoring procedure. Therefore, in order to establish a frame of reference for understanding the SCORE functions, it is necessary to begin with a detailed description of the general format, inputs, and outputs of the testing procedure. A description of the specific functions that SCORE performs appears in Section 4.0.

3.1 Procedural Description

SCORE administers a test to the user, who is seated at an interactive terminal. Two alternative answers are displayed simultaneously with each test question. One of the two answers is correct, and the other incorrect, the order being random. Figure 3-1 shows a typical display presented to the user.



ALADDIN WAS:

1. *Chinese*
2. *Persian*

Enter your answer: _____

Figure 3

A SAMPLE DISPLAY

The user must respond to each question as it is presented, citing not only the answer believed to be correct but also the degree of certainty, or probability, that the cited answer is indeed the correct one. For example, the response to the question in Figure 3-1 might be: 2 .95, indicating that the user is 95% certain that Aladdin was Persian.

Note that probability is used as the standard measure for expressing and communicating the degree of certainty. In the context of SCORE, a probability is a number between 0 and 1, inclusive, that represents the current state of the user's knowledge concerning the relative likelihoods of the two alternative answers being correct. The user's knowledge may stem from many different sources, but it must ultimately be made explicit in the form of a stated probability of being correct. Since SCORE presents two answers for consideration, the allowable range for probabilities is .5 (completely uncertain as to the correct answer) to 1.0 (absolutely certain). A probability of less than .5 would imply that the user believes that the other answer is more lik -y to be correct. For example, if the user should respond to the Aladdin question with: 1 .3, then the following inconsistency exists: the user has stated that the correct answer is that Aladdin was Chinese, but with only 30% certainty, thus at the same time implying with 70% certainty that Aladdin was Persian.

SCORE incorporates a procedure known as a proper scoring rule (designed to reduce guessing, as discussed in Chapter 30 of reference 1.2.2) which displays, for each user response, a win/lose point score that is based on the user's expressed degree of certainty. The higher the expressed probability, the more the user has to lose should the answer be incorrect. For example, should the user respond: 2 .95 to the Aladdin question, SCORE would then display: WIN 24.7 LOSE 65.2, indicating the risk involved. The user may then alter the response, if appropriate.

At the user's option, SCORE will present feedback after each question, identifying the user's response as either right or wrong and indicating the number of points won or lost. In the Aladdin example SCORE would respond: WRONG YOU LOSE 65.2 POINTS. (Surprisingly, Aladdin was Chinese.)

At the conclusion of the testing session, SCORE will compute and display two results, as described in a later section.

3.2 Testing Description

The SCORE testing format always consists of all of the following elements.

3.2.1 Available question sets - Sets of questions, each having various numbers of questions, of general interest or relating to one specific topical area.

3.2.2 Alternative responses - A set of possible responses, two for each question, one of which correctly answers the question and one of which does not.

3.2.3 Test question set - A sequence of questions selected from any set, as designated by the user, to be used during a specified testing session.

3.2.4 Correct responses - A list consisting of the correct responses to the questions.

3.2.5 User responses - A set of actual responses made by the user, one for each question, specifying which response the operator deems most likely to answer the question.

3.2.6 Probability assessments - A set of numbers between .5 and 1.0, one for each user response, specifying the user's assessed probability (degree of certainty) that the response is, in fact, correct.

3.2.7 Scoring Rule - An equation used in the evaluation of the responses, as follows:

$$\text{WIN SCORE} = 25 - 100\ (1 - P)^2$$

$$\text{LOSE SCORE} = 100\ (P)^2 - 25$$

where P is the user's probability that the answer is correct. These scores are calculated and used only to advise the user of the relative difference between the reward for a correct answer and the penalty for an incorrect one, as discussed in Section 3.1.

3.3 Results of the Model

The user's inputs are processed to produce two results: a graph showing the calibration of the probability assessments, and an overall performance measure.

3.3.1 Calibration graph - A graph that compares the actual calibration achieved by the user on the test with perfect calibration. For each of ten probability intervals, the percentage of questions answered correctly (hit rate) is graphed against the user's assessed probability. The optimal performance on the test yields a diagonal line such that the hit rate equals the assessed probabilities. Figure 3-2 displays an example of a calibration graph.

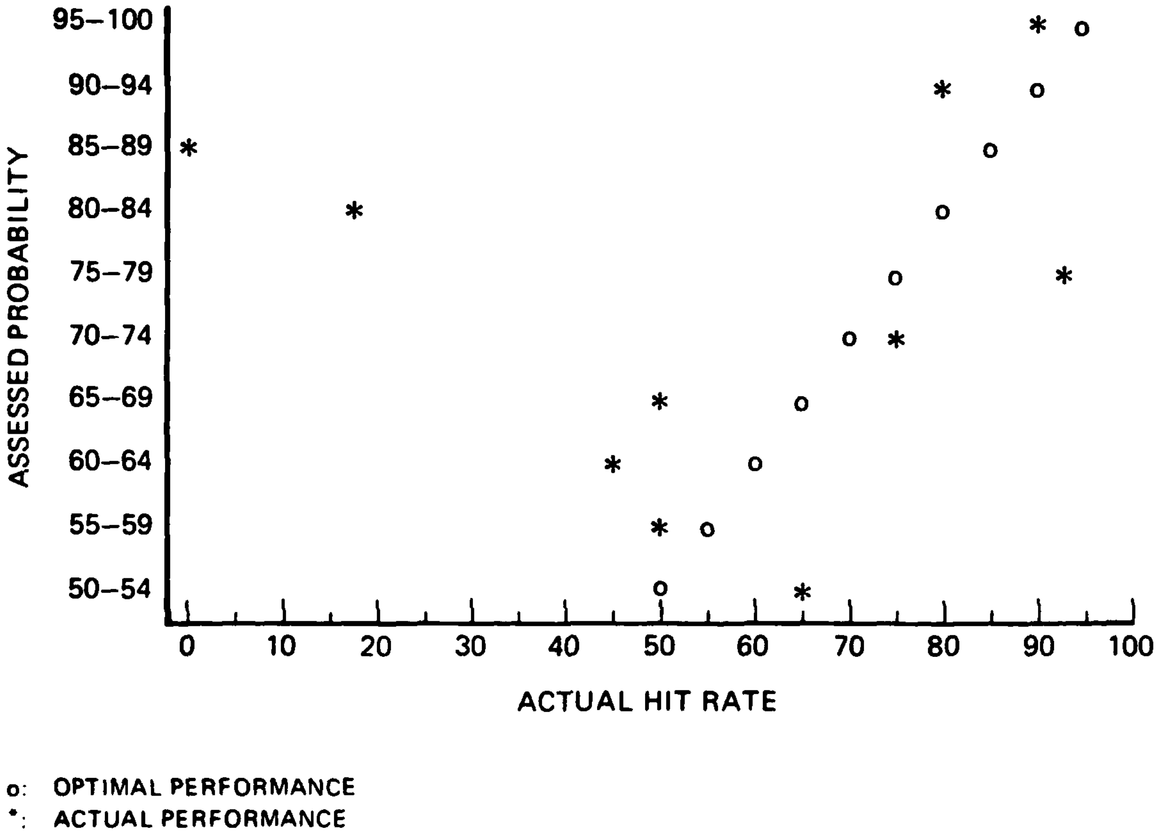


Figure 3-2

A CALIBRATION GRAPH

3.3.2 Performance measure - A user's performance on SCORE is obtained by applying a proper scoring rule, as discussed in Section 3.2.7. However, the user's performance involves two types of errors: a calibration (labeling) error and a resolution (sorting) error, as discussed in the Users Manual, reference 1.2.4. The resolution error is removed from the overall performance measure, which is computed as follows:

$$\text{OVERALL PERFORMANCE (\%)} = 100 \times \frac{1 - \text{BRIER}}{1 - \text{SORT}}$$

where BRIER and SORT are defined in the next two sections. Note that the overall performance is expressed as a percentage of the optimal performance.

BRIER - The average BRIER score is the mean square error across all questions; i.e.,

$$\text{BRIER} = \frac{\Sigma E^2}{N},$$

where Σ is the summation over all questions, N is the total number of questions answered, and E is set equal to (1-P) if the question is answered correctly and set equal to P if it is answered incorrectly. P, of course, is the user's cited probability of the answer being correct.

SORT - SORT, the average sorting error per question, is based on the user's hit rate in each of the probability intervals as shown in Figure 3-1. For example, the figure shows that the user obtained a hit rate of 65% in the assessed probability interval 50% to 54% and a hit rate of 95% in the 75% to 79% interval.

SORT is the average sorting error per question and is calculated as follows:

$$\text{SORT} = \frac{\Sigma n(H)(1-H)}{N}$$

where Σ is the summation over the ten intervals, n is the number of questions a..swered in the interval, H is the actual hit rate in the interval, and N is the total number of questions answered.

4.0 SCORE FUNCTIONS

SCORE is designed to perform the basic functions described below. A description of the detailed logical design of the SCORE functions is contained in the manual, Documentation of Decision-Aiding Software: Scoring Rule Systems Specification, reference 1.2.3.

4.1 Maintain a Library of Available Question Sets

Store various question sets, labeled by their subject matter, length, and difficulty.

4.2 Load an Existing Question Set

Display the subject matter, length, and level of difficulty of those question sets stored in the question set library, and permit the user to retrieve any desired question set or portion thereof.

4.3 Present Questions and Record User Responses

Present the questions from the selected question set, one at a time, recording the user responses to the questions. Provide feedback to the user after each question, should the user so desire. Feedback consists of the correct answer and the number of points won or lost as described in Section 3.2.7.

4.4 Display an Evaluation of the User's Performance on the Test

Display a calibration graph and overall performance measure either upon completion of the test or at any time during the test, as requested by the user by typing the letter F (FEEDBACK) instead of responding to the question.

4.5 Request Help

Permit the user to request directions describing the available options at any point during the test. The user requests help by typing the letter H (HELP), at any time.

4.6 Stop the Test Prematurely

Permit the user to stop the test at any point, obtaining final results for the completed portion of the test. This is done by typing S (STOP) instead of responding to a question.

4.7 Restart the Test

Permit the user to restart the test with a second set of questions.